The validity and reliability of a sample of ten Wattbike cycle ergometers

Key Words: Wattbike, Validity, Reliability, Ergometer, Cycling

Authors:

Barney Wainwright (corresponding author), Leeds Beckett University, Fairfax Hall, Headingley Campus, Leeds, LS6 3QS, UK, Email: barney.wainwright@leedsbeckett.ac.uk, Tel: 0113 8123532

Carlton Cooke, Leeds Trinity University, Brownberrie Lane, Horsforth, Leeds, LS18 5HD, UK, Email: c.cooke@leedstrinity.ac.uk, Tel: 0113 283 7100

John O'Hara, Leeds Beckett University, Fairfax Hall, Headingley Campus, Leeds, LS6 3QS, UK, Email: j.ohara@leedsbeckett.ac.uk, Tel: 0113 8125239

Acknowledgements: Mathew Lees, Project Assistant, Leeds Beckett University for his assistance in the collection of the data.

Funding sources: Funding for this research was provided by Wattbike Ltd, Nottingham, UK.

Disclosure statement: Other than the payment that was made by Wattbike Ltd to undertake the research study, there were no conflicts of interest between Leeds Beckett University and Wattbike Ltd arising from the research.

Word Count: 4792

The research was conducted at Leeds Beckett University

Abstract (196 words)

The purpose of the study was to assess the validity and inter-bike reliability of ten Wattbike cycle ergometers, and to assess the test-retest reliability of one Wattbike. Power outputs from 100 to 1000 W were applied using a motorised calibration rig (LODE) at cadences of 70, 90, 110 and 130 rev.min^{-1}, which created nineteen different intensities for comparison. Significant relationships ($p<0.01$, $r^2=0.99$) were found between each of the Wattbikes and the LODE. Each Wattbike was found to be valid and reliable, and had good inter-bike agreement. Within-bike mean differences ranged from 0.0 W to 8.1 W at 300 W and 3.3 W to 19.3 W at 600 W. When taking into account the manufacturers stated measurement error for the LODE (2%), the mean differences were less than 2%. Comparisons between Wattbikes at each of the nineteen intensities, gave differences from 0.6 to 25.5 W, at intensities of 152 W and 983 W respectively. There was no significant difference ($p>0.05$) between the measures of power recorded in the test-retest condition. The data suggest that the Wattbike is an accurate and reliable tool for training and performance assessments, with data between Wattbikes being able to be used interchangeably.

Introduction

The Wattbike is an air and magnetically braked cycle ergometer that was designed with British Cycling for the training and performance assessment of cyclists. Cycle ergometers, such as the Wattbike, are increasingly used across the world for assessing cycling performance and for training, and it is becoming the norm for coaches and sports science practitioners to use power output instead of heart rate to specify training intensity in cycling (Duc, Villerius, Bertucci, & Grappe, 2007). The Wattbike supports a power range from approximately 50 to 3760 Watts (W), suiting a variety of exercise applications, and has been endorsed by British Cycling for talent identification and to support their World Class Programmes (Hopker, Myers, Jobson, Bruce, & Passfield, 2010). Furthermore, numerous laboratories are known to use the Wattbike to assess performance (e.g. Driller, Argus, & Shing, 2013), and to conduct cycling-related research (e.g. Argus, Driller, Ebert, Martin, & Halson, 2013). Hence, given the varied consumer base for the Wattbike, establishing the accuracy and reliability of the power measurement is important for research (Balmer, Davison & Bird, 2000), performance assessment and training.

Due to the importance of recording power with an appropriate level of accuracy and reliability for the power meters' intended use, a number of studies have been conducted to establish the level of accuracy and reliability of commercially available cycling power meters such as the SRM (Jones & Passfield, 1998; Lawton, Martin & Lee, 1999; Gardner et al., 2004, Abbis et al., 2009), PowerTap® (Gardner et al., 2004; Bertucci, Duc, Villerius, Pernin, & Grappe, 2005), Ergomo Pro (Duc et al., 2007; Kirkland, Coleman, Wiles, & Hopker, 2008), Look Keo (Sparks, Dove, Bridge, Midgely, 2014), Polar® S710 (Millet, Tronche, Fuster, Bentley, & Candau, 2003), G-Cog (Bertucci, Crequy, & Chiementin, 2013), and power measuring cycle ergometers such as the Kingcyle (Balmer, Davison, Coleman, & Bird, 2000), Axiom Powertrain (Bertucci, Duc, Villerius, & Grappe, 2005), Velotron (Abbis, Quod, Levin, Martin, & Laursen, 2009), Wattbike (Hopker et al., 2010) and a new design of ergometer (Bertucci, Grappe, & Crequy, 2011).

Motorised calibration

Although in many studies the SRM powermeter has been used as the criterion measure, an alternative, appropriate method reported in the literature to assess the validity of power measurement systems and ergometers is through the use of motorised calibration rigs (Wilmore et al., 1982; Russell & Dale, 1986; Maxwell et al, 1998; Jones & Passfield, 2000; Lawton, Martin, & Lee, 1999; Abbis et al., 2009). Although various designs have been employed, the most common type of motorised calibration rig used now incorporates a speed-controlled motor to apply a torque to the bicycle pedal or bottom bracket via a crankshaft. As such it is essentially a 'torque reaction measuring device', where power output is calculated as the product of torque and angular velocity. The torque is measured using a high quality load cell placed at a known distance from the rotational axis of the crankshaft, and the angular velocity is measured by a tachometer. The manufacture and specifications of such motorised calibration rigs has been described by Woods, Day, Withers, Ilsley, & Maxwell (1994) and Drouet, Champoux, & Bergeron (2008), and given that they function on a first principles basis, are accurate and reliable if quality components are used to measure the applied force and the angular velocity of the rotating crankshaft. Estimated errors as low as 0.3% up to 353 W, with a variation of 0.6 to 3.2%, (Woods et al., 1994) and 0.9% between 50 and 600 W (Drouet et al., 2008) have been reported, but not externally validated. Calibrating the load cell prior to use and ensuring a constant environmental temperature during use are essential to maintain high levels of reliability and accuracy.

Existing studies describing the accuracy and reliability of the Wattbike

Hopker et al. (2010) assessed the validity and reliability of a single Wattbike by comparing it to an SRM powermeter (Science model) that was fitted to the Wattbike in place of its' own chainset and cranks. The study was conducted in two parts. In the first part a comparison was made between the power recorded by the Wattbike and the SRM while a motorised calibration rig applied a power input between 50 W and 1250 W using cadences of 70 and 90 rev.min^{-1}. In the second part, power outputs from ten trained and ten untrained cyclists at 4 submaximal work rates, and a 5 minute performance trial were compared. According to their report the SRM was calibrated at the start of the study with the zero being reset prior to each trial. In the trials that used a motorised calibration rig significant differences ($p<0.05$) were found between SRM and Wattbike at both 70 and 90 rev.min^{-1} at all 38 power outputs tested (except 100, 550 and 600 W in the 90 rev.min^{-1} trial), although strong correlations between SRM and Wattbike power were found at both cadences ($r=0.99$). These differences resulted in an agreement of ± 1.7% and ± 1.4% at the 70 and 90 rev.min^{-1} power outputs respectively. In the steady state trials there were significant differences that ranged from -7% (300 W, trained group) to 16% (50 W, untrained group), although the mean difference across all power outputs was -0.4%. In the performance trials significant differences were found in both the untrained group ($p<0.01$, 234 W vs. 239 W respectively, 95% limits of agreement -21 to 11 W) and the trained group ($p=0.03$, 310 W vs. 339 W respectively, 95% limits of agreement -4 to 62 W). The Wattbike recorded higher levels of variability in the repeatability trials than the SRM (coefficient of variation of 6.7% and 2.6% for the Wattbike in the untrained and trained groups vs 2.2% and 1.1% for the SRM). As a result of the differences found across the whole study, the authors explained that although the overall mean error of <2% would be sufficiently accurate in most situations, some of the absolute differences, which were in the region of 23 W, may be too large in an elite population where greater precision is required.

Although Hopker et al. (2010) found differences between the Wattbike and the SRM, it is the opinion of the manufacturers of the Wattbike (personal communication) that the replacement of the existing non-standard crankset with the SRM, will have invalidated the ability of the Wattbike to accurately measure power. The Wattbike relies on fine tolerances of chain tension and chain alignment for reliable measures of force, and the change of crankset, even if it had suitable dimensions, would have required a recalibration before use, something that was not performed in the study. Therefore, the outcomes of the Hopker et al. (2010) study should be considered with caution.

Aims and objectives of the study

The purpose of this study was to establish the validity and reliability of a pool of ten Wattbikes, and to assess the test-retest reliability of a single Wattbike.

Methods

Testing was carried out using ten new Wattbike cycle ergometers (Wattbike Ltd, Nottingham, UK), which had been calibrated during manufacture. These were selected at random from the distribution warehouse by a member of the research team the week prior to testing, and then transported to the testing laboratory in Leeds.

Prior to each test the left crank arm was removed from each Wattbike to allow the motorised calibration rig (Lode Calibrator 2000, Groningen, Netherlands) (LODE) to be attached directly to the bottom bracket. The LODE measured the rotational torque applied via a load cell that was pre-calibrated using seven calibration weights (1 to 7 kg, in 1 kg increments) the day immediately prior to the data collection, and the rotational velocity was measured by a tachometer. The LODE had a manufacturers stated error of ±2%, a cadence accuracy of 0.1 rev.min^{-1}, and a torque accuracy of 0.04 Nm. The temperature controlled laboratory was maintained at 20˚C throughout the experimental trials to help maintain the reliability of the load cells in the LODE and Wattbikes. The barometric pressure was recorded during the same period and ranged from 1012.6 hPa to 1013.9 hPa. Prior to each trial the LODE offset and the Wattbike Zero offset were reset. A single calibration weight (4 kg) was used as a check of drift of the LODE load cells' calibration immediately prior to testing the first Wattbike, and at the end of the data collection. The Wattbike calculates power output by measuring the load applied to a load cell as a result of chain tension at sampling rate of 100 Hz using the formula:

$$P = F. \left(\frac{2.\pi.l_C}{t}\right)$$

Where P = power output per revolution (W), F = average force per crank revolution (N), l_C = crank length (0.17 m), and t = time taken to complete the crank revolution (s). Cadence is measured twice per pedal revolution. Each Wattbike is calibrated via a motorised calibration rig in the manufacturing plant, and the power calculated using individual calibration coefficients that are stored within each bikes firmware.

Once the LODE was attached to a Wattbike it drove the ergometer at cadences of 70, 90, 110, and 130 rev.min^{-1}. At each cadence the power outputs were achieved by manually adjusting the resistance settings on the Wattbike via the air resistance lever arm to positions of 2, 4, 6, 8 or 10. Given that the actual resistance applied in each position is dependent upon the air density at the time of use, determined primarily by the ambient pressure and temperature, the lever arm positions were used as initial targets from which smaller adjustments were made to fine-tune the power output. This method was used to attain approximate power outputs of: 90 W, 120 W, 150 W, 180 W, and 200 W at 70 rev.min^{-1}, 160 W, 220 W, 300 W, 350 W and 400 W at 90 rev.min^{-1}, 260 W, 380 W, 520 W, 620 W, and 700 W at 110 rev.min^{-1}, and 400 W, 600 W, 830 W and 980 W at 130 rev.min^{-1}. Five stages per cadence were used, with the exception of 130 rev.min^{-1}, where four stages were used, as 990 W was the limit of the LODE. The magnetic brake was not applied during the trials. As a result of the practicalities of the methods employed, power outputs close to but not exactly at the target power outputs were recorded. This resulted in a time efficient and consistent measurement protocol yet still allowed a methodologically sound comparison of power outputs between the Wattbikes and the LODE. Each power output stage for a given cadence lasted 1 minute. The initial 30 seconds was used to adjust the resistance lever arm to attain the approximate target power output and allow the LODE to stabilise, whilst the last 30 seconds of data were concurrently recorded using both the LODE and the Wattbike. Thereafter, there was an increase in resistance with or without an associated change of cadence. Power output readings from the LODE were recorded at 5 second (70 rev.min^{-1}) or 3 second (90, 110, and 130 rev.min^{-1}) intervals depending on the cadence, and used to determine the average power input during each stage. The corresponding data was downloaded using the Wattbike Expert software package (Wattbike Ltd, Nottingham, UK), which recorded actual cadence, force, torque, and power output data from the Wattbike for every pedal stroke. The above procedures were followed for each Wattbike tested, with the exception of one Wattbike. A randomly selected Wattbike was also used to repeat the above protocol on two occasions to assess the Wattbike's repeatability. The proposed study was approved by the University's Ethics Committee, and carried out in accordance with the University's health and safety guidelines.

Statistical Analysis

Mean power output values were calculated for each 30 second period from the LODE and the Wattbike for comparative purposes. The difference (residual) in power output of the Wattbike compared to the calibrator was computed by subtracting the Wattbike power output (30 second average) from that recorded by the LODE. Linear regression was used to determine the relationship between the LODE power input and the power output measured by the Wattbike. The bias and 95% limits of agreement between the Wattbike and LODE were calculated using the methods of Bland & Altman (1986) for between bike comparisons, and the revised Bland & Altman (1999) method for heteroscedastic data when comparing within and across bikes. The revised method resulted in a linear regression model that described the relationship between the measurement value and the magnitude of the bias. A one-way ANOVA was used to assess the data from the repeated test on one of the Wattbikes. The IBM SPSS version 20 was used to carry out the statistical analysis. The minimum level of level of significance accepted was $P < 0.05$.

Results

The differences between Wattbike power and LODE power were investigated using 190 pairs of data from 10 Wattbikes. Seven pairs of data (3.7% of the total data recorded) were removed due to an irregular propagation of the force signal in the LODE that occurred for either a 3 or 5 second period within the 30 second data collection period. This error was found retrospectively following careful interrogation of the raw signals. A conservative approach was taken by removing these data sets from the analysis, rather than only removing the portion of affected data. This avoided any bias that might have been created by attempting to use the remaining data in the analysis. A further 19 pairs of data were recorded to assess the reliability of one of the Wattbikes. No calibration drift was found in the load cell of the LODE over the duration of the testing, with the 4 kg calibration weight reporting the same calibration value before and after the trials.

Validity

The power recorded by each of the Wattbikes was compared to the power input from the LODE. The relationship and variation between the LODE power applied and the Wattbike power measured for one of the Wattbikes can be seen in Figure 1.

Figure 1 near here

The regression models for each of the bikes (Table 1) show that there was little variation in the model coefficients between each bike (Table 1). These results show that a change in power in one bike will be matched by a very similar, or in some cases identical, change in power in another Wattbike.

****Table 1 near here****

****Figure 2 near here****

It can be seen in Figure 2 that in general the differences were larger at the higher power outputs and there was a general trend that the majority of the Wattbikes read higher than the LODE. The relationship between the measurement value and the differences was described by the regression model:

$$d = 0.029.\ x - 4.624$$

Where d is the difference between the two measures, and x the LODE power output. This model can be used to determine the bias and limits of agreement at any value within the experimental range. Examples of the values of bias and 95% confidence intervals using the regression model can be seen in Table 2. As illustrated by the data, the bias increases in magnitude as the level of power output increases. When considering the full range of power outputs the mean bias is 11.2 W (1.5%), although this varies from -1.8 W (-1.8%) at 100 W to 24.1 W (2.4%) at 1000 W.

****Table 2 near here****

The largest difference between the Wattbike and LODE was 30.5 W (3.8%) at 130 revs.min^{-1} for Bike 9 (Table 3). In percentage terms the largest difference was -7.1% (-6.2 W) for Bike 5 (Table 3), which occurred at the workload with the lowest power (approximately 85 W).

****Table 3 near here****

The LODE has a manufacturers stated error of ±2%, and this should be taken into account when comparing it to the Wattbike. Figure 3 shows the same plot of the residuals between the LODE and Wattbike, but with 2 lines illustrating both the ±2% LODE error.

Figure 3 near here

To account for the LODE manufacturers stated error, the residuals can be calculated as percentages. Figure 4 shows the residual as a percentage between the Wattbike and LODE power +2% or -2%, which ever is the smaller difference. All data points other than two from Bike 5 are within the 2% boundaries.

****Figure 4 near here***

<u>Within Bike variations</u>

The variation between the power input from the LODE and the recorded power from the Wattbike was assessed for each bike across the measurement range. The summary data can be seen in Table 4, with the breakdown of the differences for each bike in Table 3.

Table 4 near here

The mean value (W) indicates the bias for each bike, which was positive in all cases, and the limits of agreement identify the range of power outputs within which 95% of the differences between the Wattbike and LODE measurements would lie for each of the bikes. All bikes demonstrate a very similar range of limits of agreement, with the exception of bikes 7 and 10. They had little bias and a small range in their limits of agreements (-7.5 to 9.7 W and -6.9 to 7.5 W respectively), suggesting that these two bikes record very similar values to the LODE across the range of power applied. Bike 5 was the worst performing bike with the largest SD (11.5 W) and the largest limits of agreement (-19.0 to 26.2 W). When considering the data from the whole group of bikes, it demonstrates very good levels of within bike variation.

Between bike variations

Of importance is that each Wattbike has a similar level of accuracy when training or testing at similar workloads on different bikes. Table 5 shows the magnitude of variation across the Wattbikes at each cadence and resistance setting.

Table 5 near here

The variations in the differences in the power between the Wattbike and the LODE between the bikes at each of the intensities were small (absolute differences ranging from 0.6 W at an intensity workload of 152 W, to 25.5 W at an intensity of 983 W), demonstrating a very high reproducibility of the measurement of power between bikes. The mean differences increase with intensity, and the largest difference of 25.5 W represents 2.6% of the mean power of 983 W, at that intensity.

Single bike repeat tests

The difference between residuals and limits of agreement for Bike 6 and a repeat test of Bike 6 can be seen in Figure 5. In each test the Wattbike was compared against the LODE, and both tests took place on the same day, separated by a period of eight hours.

Figure 5 near here.

Figure 5 shows high test-rest reliability between the two tests, where repeatability in the residuals is apparent at the same experimental power outputs. The regression models demonstrate similar gradients, only differing in intercept. The upper and lower 95% limits of agreement show a large degree of overlap between the two sets of data, illustrating the expected similarity in the recorded values. When comparing the results at 300 W and 600 W, the differences between the residuals were 1.5 W and 1.7 W respectively. A one-way ANOVA found no significant difference ($p<0.05$) between the values recorded by Bike 6 in test 1 and test 2.

Discussion

The purpose of the study was to establish the validity, reliability and repeatability of a random selection of ten Wattbikes by comparing the recorded power to the power applied by a motorised calibration rig (LODE). There were significant relationships ($p<0.01$, $r^2=0.99$) between the Wattbike and LODE power between 100W and 1000W in each of the Wattbikes. The mean differences between each of the Wattbikes and the LODE were less than 2% when considering the manufacturers stated error (2%) in the application of power by the LODE. This is in contrast to the results of Hopker et al. (2010) who found significant differences between the Wattbike and SRM in 35 of the 38 power outputs used, with differences ranging from -7% at 300 W to 16% at 50 W. The Wattbikes, in the present study, were found to be both valid and reliable, with the mean differences ranging from 0.0% (0 W) to 2.8% (8.4 W) at 300 W, and 0.8% (4.9 W) to 3.2% (19.3 W) at 600 W without accounting for the manufacturers stated error (2%) in the LODE. When considering the data from all of the bikes the magnitude of the error increased with larger power outputs. For example at 200 W the bias was 1.1 W (0.6%) with lower and upper 95% limits of agreement ranging from -7.5 W (-3.8%) to 9.8 W (4.9%), and at 1000 W the bias was 24.1 W (2.4%) with the lower and upper 95% limits of agreement ranging from 15.4 W (1.5%) to 32.7 W (3.3%).

The results for the Wattbike can be compared to other studies that have compared air-braked ergometers to a motorised calibration rig. Abbis et al. (2009) investigating the

Velotron cycle ergometer found errors of 0.80% and -0.34% during constant paced intensities of 250 and 414 W respectively, and mean errors of 3.0% (95% confidence intervals of 1.6 – 4.5%) for average power and -55.8% (95% confidence intervals of -55.9 – -55.7%) for peak power during three 35 second high intensity intervals. In incremental trials (180 to 1320 W) they recorded an average of 1.9% error (95% confidence intervals of -2.2 – 6.0%), with larger errors (42% and 19%) at high workloads (>1200 W). Maxwell et al. (1998) assessed the accuracy of five air-braked scientific grade Repco cycle ergometers, finding mean errors of 0.0% to 1.6%, and -0.4% and 1.4% across a range of power outputs from 274 W to 1120 W. The magnitude of error in individual bikes ranged from -3.3% to 1.5% in peak power, and -3.3% to 2.0% in mean power.

When comparing across Wattbikes at each cadence and resistance level the mean differences range from a 0.6 W difference at 70 $rev.min^{-1}$ (mean power input of 152 W) to a 25.5 W difference at 130 $rev.min^{-1}$ (mean power input of 983 W). The Wattbike was found to have high levels of repeatability during the test-retest protocol, with the individual regression models between the differences and the power outputs being almost identical. When comparing the results at 300 W and 600 W, the differences between the residuals were 1.5 W and 1.7 W respectively. The differences between the LODE and the Wattbike were 5.2 W and 6.7 W at 300 W, and 13.6 W and 15.3 W at 600 W in the first trial second trial respectively. No significant differences found between the LODE and Wattbike in the two repeated trials. The results should give confidence to the user as they show that all of the Wattbikes tested were accurate in their measurement of power output, and the results from one bike to another are very similar in terms of the magnitudes of difference reported in comparison to the motorised calibration rig. This is of particular importance for talent identification assessments, team or squad training or in the physiological assessment of athletes, where often tests and trials will take place in different locations. These results provide sufficient confidence that the results from tests carried out on different Wattbikes can be directly compared, although it is still advisable wherever possible to conduct longitudinal monitoring of cyclists on the same bike when the changes between tests are expected to be small. The results also show that the day-to-day comparison of data obtained from the use of a single Wattbike are reliable, with a very small (0.6 W) difference between the repeated trials reported. This finding allows the user to be confident that their day-to-day results are consistently measured and any changes in power observed are real and not as a result of any unreliability on the part of the Wattbike.

When making assessments of comparative data it is very important to put them in the context of the accuracy of the criterion measure used. As discussed earlier, appropriate and precise calibration routines are essential, and must be reported for a true comparative assessment to be made and understood by the reader as well as the research team. Ideally a post-trial check of the stability of the pre-trial calibration should be undertaken and reported. In this study, it was not possible, nor necessary, to validate the criterion measure used. However, a careful seven point pre-trial calibration was made of the system the day prior to testing, with a single calibration weight used to check the calibration for drift before and after the trials that took place over a 10 hour period. Even though a careful calibration and data collection process was used, the criterion measure had a manufacturers stated error of ±2%, and this should be taken into account when the data from the LODE is compared to the Wattbike. The assessment of other studies investigating the validity of other powermeters should take these points into consideration, especially in those studies that have not reported the procedures in sufficient detail to provide confidence that they have ensured optimal accuracy of the criterion measure (Wooles, Robinson, & Keen, 2005).

The current study used a different approach to measure the validity of the Wattbike to that of Hopker et al. (2010). They used a scientific model SRM powermeter fitted to the Wattbike, to record the human and mechanical applied power, to act as the criterion measure. However, the manufacturers of the Wattbike claim that this process will have invalidated the Wattbikes ability to measure power accurately and reliably. Although they reported a strong correlation between the SRM and Wattbike (r=0.99), as well as significant differences, the data should be interpreted with some caution given the methods that were employed, as described in the introduction. Therefore, making a direct comparison of the results from the Hopker et al. (2010) study and the present study is not appropriate. In the present study the use of a motorised calibration rig became practically most appropriate due to the challenges of placing an SRM powermeter on the Watttbike and ensuring that the Wattbike would operate ‘as manufactured’. Repeating the installation of a scientific model SRM and the ‘in-factory’ recalibration process for nine other Wattbikes, if possible at all, would have become logistically prohibitive and was not possible in the present study. The more practical, and equally effective methodology used in this study does mean however that the use of human participants to increase the ecological validity of the applied power was not possible, but from a calibration perspective this approach excludes one source of variation from the data.

There were some limitations in the current study that may be possible to overcome in future studies. Due to the time constraints of the data collection period and the magnitude of the experimental protocol, data was collected over a 30 second period for each experimental intensity after a 30 second period for stabilisation at the new intensity. It could be argued that a longer period should be used to collect the data, but given previous experience in using the LODE it was decided that this was not necessary. None of the collected experimental data suggested that a longer collection period was required. In addition, the reliability of the Wattbikes to measure power over longer periods of time (i.e. greater than 20 minutes) was not assessed. While there was nothing to suggest that the validity might be affected during longer bouts of use, this could be assessed in future studies.

The identification of the ‘Gold Standard’ measure for the validation of ergometers and powermeters remains to be resolved. While previous research supports the notion that motorised calibration rigs should be the reference measurement (Maxwell et al., 1998; Woods et al., 1994; Gardner et al., 2004), most commercially available motorised calibration rigs that are regularly used in sport science laboratories state an accuracy in the region of 2%, and not necessarily over the range of power outputs found when measuring human performance. While some crank-based powermeters may offer an alternative method, the choice of powermeter, and the calibration process employed and reported, must be given careful consideration. When considering the validity of ergometers or powermeters, careful assessment should be given to the methodology used.

Conclusions

The current study has assessed ten randomly selected new Wattbikes across a range of power outputs (100 W to 1000 W) using sport specific cadences (70, 90, 110, 130 $rev.min^{-1}$). While accounting for the stated error in the motorised calibration rig (2%) mean differences of less than 2% were found across the ten Wattbikes. In addition, the Wattbike has been found to be highly reliable both between bikes (0.6 and 25.5 W differences at 100 W and 1000 W respectively) and within repeated measures on the same bike (measurement differences of 1.5 W and 1.7 W between trials at 300 W and 600 W respectively). These results provide the user with confidence that the Wattbike is an accurate and reliable tool for training and performance assessments, with data between Wattbikes being able to be used interchangeably.

References

Abbiss C.R., Quod M.J., Levin G., Martin D.T., Laursen P.B. (2009). Accuracy of the Velotron Ergometer and SRM Power Meter. *International Journal of Sports Medicine*, 30(2), 107-112. doi: 10.1055/s-0028-1103285

Argus C.K., Driller M.W., Ebert T.R., Martin D.T., Halson S.L. (2013). The Effects of 4 Different Recovery Strategies on Repeat Sprint-Cycling Performance. *International Journal of Sports Physiology and Performance*, 8, 542-548.

Balmer J., Davison R.C.R., Bird S.R. (2000). Reliability of an air-braked ergometer to record peak power during a maximal cycling test. *Medicine and Science in Sports and Exercise*, 32, 1790-1793.

Balmer J., Davison R.C., Coleman D.A., Bird S.R. (2000). The validity of power output recorded during exercise performance tests using a Kingcycle air-braked cycle ergometer when compared with an SRM powermeter. *International Journal of Sports Medicine*, 21, 195-199. doi: 10.1055/s-002-229

Bertucci W., Duc S., Villerius V., Pernin J.N., Grappe F. (2005). Validity and reliability of the Powertap mobile cycling powermeter when compared with the SRM device. *International Journal of Sports Medicine*, 26, 868-873. doi: 10.1055/s-2005-837463

Bertucci W., Duc S., Villerius V., Grappe F. (2005). The Axiom cycling ergometer is not a valid device when compared with the SRM. *International Journal of Sports Medicine,* 26, 59-65. doi: 10.1055/s-2004-817855

Bertucci W., Crequy S., Chiementin X. (2013). Validity and reliability of the G-Cog BMX powermeter. *International Journal of Sports Medicine*, 34, 538-543. doi: 10.1055/s-0031-1301319

Bertucci W., Grappe F., Crequy S. (2011). Original characteristics of a new cycle ergometer. *Sports Engineering*, 13, 171-179. doi: 10.1007/s12283-011-0063-6

Bland J.M., Altman D.G. (1986). Statistical methods for assessing agreement between 2 methods of clinical measurement. *The Lancet*, 327(8476), 307-310. doi: 10.1016/s0140-6736(86)90837-8

Bland J.M., Altman D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8, 135-160. doi: 10.1177/096228029900800204

Driller M.W., Argus C.K., Shing C.M. (2013). The Reliability of a 30-s Sprint Test on the Wattbike Cycle Ergometer. *International Journal of Sports Physiology and Performance*, 8, 379-383.

Drouet J., Champoux Y., Bergeron F. (2008). A user-friendly calibration system for bicycle ergometers, home trainers and bicycle power monitoring devices. *Sports Engineering*, 11, 15-22. doi: 10.1007/s12283-008-0003-2

Duc S., Villerius V., Bertucci W., Grappe F. (2007) Validity and Reliability of the Ergomo Pro Power Meter Compared with the SRM and Powertap Power Meters. *International Journal of Sports Physiology and Performance*, 2, 270-281.

Gardner A.S., Stephens S., Martin D.T., Lawton E., Lee H., Jenkins D. (2004). Accuracy of SRM and Power Tap power monitoring systems for bicycling. *Medicine and Science in Sports and Exercise,* 36, 1252-1258.

Hopker J., Myers S., Jobson S.A., Bruce W., Passfield L. (2010). Validity and Reliability of the Wattbike Cycle Ergometer. *International Journal of Sports Medicine*, 31, 731-736. doi: 10.1055/s-0030-1261968

Jones S.M., Passfield L. (1998). The dynamic calibration of bicycle power measuring cranks. In S.J. Haake (Ed.): The Engineering of Sport, (pp. 265-274). Oxford: Blackwell Science Ltd.

Kirkland A., Coleman D., Wiles J.D., Hopker J. (2008). Validity and reliability of the Ergomopro powermeter. *International Journal of Sports Medicine,* 29, 913-916. doi: 10.1055/s-2008-1038621

Lawton E.W., Martin D.T., Lee H. (1999). Validation of SRM power cranks using dynamic calibration. In: Proceedings of the 5th IOC World Congress on Sport Sciences, Oct 31 – Nov 5, 1999. Sydney: International Olympic Committee.

Maxwell B.F. Withers R.T., Ilsley A.H., Wakim M.J., Woods G.F., Day L. (1998). Dynamic calibration of mechanicaly, air- and electromagnetically braked cycle ergometers. *European Journal of Applied Physiology*, 78, 346-352.

Millet G.P., Tronche C., Fuster N., Bentley D.J., Candau R. (2003). Validity and reliability of the Polar®S710 mobile cycling power meter. *International Journal of Sports Medicine*, 24, 156-161. doi: 10.1055/s-002-2566

Russell J.C., Dale J.B. (1986). Dynamic torquemeter calibration of bicycle ergometers. *Journal of Applied Physiology*, 61, 1217-1220.

Sparks S.A., Dove B., Bridge C.A., Midgely A.W., McNaughton L.R. (2014). Validity and reliability of the Look Keo Power Pedal system for measuring power output during incremental and repeated sprint cycling. (2014). *International Journal of Sports Physiology and Performance*, 10, 39-45. doi: 10.1123/ijspp.2013-0317.

Wilmore J.H., Constable S.H., Stanforth P.R., Buno M.J., Tsao Y.W., Roby F.B., Lowdon B.J., Ratliff R.A. (1982). Mechanical and physiological calibration of 4 cycle ergometers. *Medicine and Science in Sports and Exercise*, 14, 322 – 325.

Woods G.F., Day L., Withers R.T., Ilsley A.H., Maxwell B.F. (1994). The dynamic calibration of cycle ergometers. *International Journal of Sports Medicine*, 15, 168-171. doi: 10.1055/s-002-8994

Wooles A.L., Robinson A.J., Keen P.S. (2005). A static method for obtaining a calibration factor for SRM bicycle power cranks. *Sports Engineering*, 8, 137-144.

Tables

Table 1. Regression models, Coefficient of Determination (r^2) and level of significance of the relationship (p-value) for each bike between Wattbike power and LODE power.

	n	r^2	p-value	Model Coefficient	Model Constant
Bike 1	18	0.99	0.00	0.972	0.120
Bike 2	19	0.99	0.00	0.973	4.056
Bike 3	18	0.99	0.00	0.972	3.250
Bike 4	18	0.99	0.00	0.967	5.116
Bike 5	19	0.99	0.00	0.960	11.965
Bike 6	19	0.99	0.00	0.973	3.213
Bike 7	16	0.99	0.00	0.983	5.452
Bike 8	19	0.99	0.00	0.976	2.393
Bike 9	19	0.99	0.00	0.964	3.032
Bike 10	18	0.99	0.00	0.987	4.384

Table 2. The bias, and lower and upper 95% limits of agreement (LoA) expressed in watts (W) and %, between the LODE power and Wattbike power calculated from the regression model for a range of power outputs.

Power	Bias		Lower LoA		Upper LoA	
(W)	(W)	(%)	(W)	(%)	(W)	(%)
100	-1.8	-1.8	-10.4	-10.4	6.9	6.9
200	1.1	0.6	-7.5	-3.8	9.8	4.9
300	4.0	1.3	-4.6	-1.5	12.6	4.2
400	6.9	1.7	-1.8	-0.4	15.5	3.9
500	9.7	1.9	1.1	0.2	18.4	3.7
600	12.6	2.1	4.0	0.7	21.2	3.5
700	15.5	2.2	6.8	1.0	24.1	3.4
800	18.3	2.3	9.7	1.2	27.0	3.4
900	21.2	2.4	12.6	1.4	29.8	3.3
1000	24.1	2.4	15.4	1.5	32.7	3.3
Mean	11.2	1.5	2.5	-1.0	19.8	4.0

Table 3. Absolute (W) and relative (%) differences between the LODE power and the power recorded by the Wattbike. Positive values indicate that the Wattbike power recorded was greater than the LODE.

Cadence	Lever position	Bike 1		Bike 2		Bike 3		Bike 4		Bike 5		Bike 6		Bike 7		Bike 8		Bike 9		Bike 10	
(rev.min^{-1})		(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)
70	2	1.1	1.4	-1.9	-2.2	-0.7	-0.8	-2.7	-3.2	-6.2	-7.1	-0.6	-3.0	-3.6	-4.0	-0.2	-0.2	-0.1	-0.1	-3.0	-3.3
70	4	2.0	1.8	-0.4	-0.4	-0.3	-0.3	-1.5	-1.3	-6.0	-5.0	0.2	-2.1	-2.6	-2.1	1.2	1.0	1.0	0.9	-2.1	-1.8
70	6	4.0	2.7	1.3	0.9	1.6	1.0	0.6	0.4	-5.8	-3.7	1.7	-1.0	-1.5	-1.0	2.4	1.6	2.8	1.9	-1.0	-0.6
70	8	5.7	3.3	2.6	1.5	2.6	1.4	2.0	1.1	-5.6	-3.1	2.3	-0.6	-1.4	-0.7	3.0	1.6	4.2	2.4	-0.6	-0.3
70	10	6.4	3.3	3.7	1.9	3.5	1.8	3.1	1.6	-5.3	-2.6	3.2	0.3	-0.6	-0.3	3.5	1.7	5.6	2.8	0.3	0.1
90	2	3.3	2.2	-0.8	-0.5	1.5	0.9	-1.2	-0.8	-7.9	-5.1	0.2	-3.2	-3.3	-2.0	0.8	0.5	2.1	1.4	-3.2	-2.0
90	4	4.4	2.1	1.6	0.7	3.7	1.6	2.0	0.9	-7.3	-3.2	2.1	-1.1			2.6	1.2	4.9	2.3	-1.1	-0.5
90	6	8.5	2.9	5.6	1.9	5.2	1.7			-7.8	-2.6	5.8	-0.4	0.5	0.2	4.9	1.7	8.5	2.9	-0.4	-0.1
90	8	11.4	3.2	7.0	2.0	7.0	2.0	7.0	2.0	-5.7	-1.6	7.5	0.8			6.5	1.8	10.8	3.1	0.8	0.2
90	10	12.6	3.2	7.2	1.8	9.2	2.3	7.5	1.9	-5.8	-1.5	9.6	2.0	3.9	1.0	8.4	2.1	13.2	3.3	2.0	0.5
100	2	6.0	2.5	0.1	0.0	3.3	1.3	2.0	0.8	7.6	3.2	1.8	-4.1	-2.9	-1.1	2.7	1.0	4.4	1.7	-4.1	-1.5
100	4	10.2	2.8	6.0	1.6	7.0	1.8	8.4	2.2	10.2	2.8	7.7	-1.8	-0.3	-0.1	5.9	1.5	10.0	2.7	-1.8	-0.5
100	6	14.5	2.9	9.4	1.8	12.3	2.4	13.9	2.7	11.6	2.3	14.4	3.7	3.4	0.6	11.1	2.1	17.3	3.4	3.7	0.7
100	8	19.9	3.3	14.5	2.4	16.4	2.7	17.5	2.9	15.5	2.6	17.1	7.1	6.6	1.1	14.5	2.3	22.2	3.6	7.1	1.1
100	10	20.6	3.0	17.5	2.5	19.7	2.8	21.8	3.1	17.5	2.5	19.2	8.1	8.1	1.1	15.5	2.2	24.5	3.5	8.1	1.1
130	2	9.7	2.5	1.9	0.5	6.0	1.5	3.4	0.8	9.3	2.5	3.7	-5.3	-2.7	-0.7	3.9	1.0	7.1	1.8	-5.3	-1.3
130	4			9.1	1.5	12.6	2.1	12.6	2.1	13.3	2.3	11.4	0.7	3.0	0.5	10.6	1.7	16.9	2.9	0.7	0.1
130	6	22.5	2.7	19.4	2.3	19.40	2.3	21.8	2.7	21.2	2.6	19.2	5.0	10.5	1.3	17.6	2.1	30.5	3.8	5.0	0.6
130	8	25.2	2.6	24.4	2.5			26.8	2.7	26.1	2.7	23.3	2.4			22.6	2.3	30.1	3.1		

Missing values: data were removed due to an irregular propagation of the force signal in the LODE.

Table 4. Bias expressed and lower and upper 95% limits of agreement (LoA) of the differences between the LODE power and Wattbike power for each Wattbike at 300 W and 600 W. Differences expressed in watts (W) and percentages (%).

	300 W						600 W					
	Bias		Lower LoA		Upper LoA		Bias		Lower LoA		Upper LoA	
	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)	(W)	(%)
Bike 1	8.4	2.8	5.9	2.0	10.9	3.6	16.9	2.8	14.4	2.4	19.4	3.2
Bike 2	4.2	1.4	0.4	0.1	8.1	2.7	12.6	2.1	8.7	1.5	16.5	2.7
Bike 3	5.4	1.8	3.1	1.0	7.7	2.6	14.1	2.4	11.8	2.0	16.5	2.7
Bike 4	4.8	1.6	0.5	0.2	9.2	3.1	15.0	2.5	10.6	1.8	19.3	3.2
Bike 5	0.0	0.0	-11.5	-3.8	11.5	3.8	12.3	2.0	0.8	0.1	23.8	4.0
Bike 6	5.2	1.7	1.5	0.5	8.8	2.9	13.6	2.3	10.0	1.7	17.3	2.9
Bike 7	-0.3	-0.1	-6.5	-2.2	5.9	2.0	4.9	0.8	-1.3	-0.2	11.1	1.9
Bike 8	4.9	1.6	1.5	0.5	8.3	2.8	12.3	2.0	8.9	1.5	15.6	2.6
Bike 9	8.1	2.7	2.6	0.9	13.6	4.5	19.3	3.2	13.8	2.3	24.8	4.1
Bike 10	-0.6	-0.2	-5.2	-1.7	4.1	1.4	3.3	0.5	-1.4	-0.2	7.9	1.3
Mean	4.0	1.3	-0.8	-0.3	8.8	2.9	12.4	2.1	7.6	1.3	17.2	2.9

Table 5. The variation across the ten Wattbikes at each of the workloads. Values shown are the mean differences, Standard Deviation (SD) and 95% limits of agreement (LoA) of the differences measured between the LODE and Wattbikes. The individual data for each bike can be found in Table 3.

Cadence ($rev.min^{-1}$)	Lever position	Mean LODE Power Input (W)	Mean Difference (W)	SD (W)	Lower LoA (W)	Upper LoA (W)
70	2	87	-1.8	2.2	-6.0	2.4
70	4	118	-0.9	2.3	-5.5	3.7
70	6	152	0.6	2.8	-4.8	6.1
70	8	179	1.5	3.2	-4.8	7.8
70	10	200	2.4	3.4	-4.3	9.0
90	2	157	-0.9	3.3	-7.3	5.6
90	4	221	1.5	3.7	-5.8	8.7
90	6	300	3.4	5.2	-6.8	13.6
90	8	352	5.8	5.3	-4.5	16.1
90	10	401	6.8	5.6	-4.2	17.8
110	2	257	2.1	3.7	-5.1	9.2
110	4	379	6.3	4.2	-2.0	14.6
110	6	519	11.2	4.6	2.2	20.1
110	8	618	15.1	5.0	5.4	24.8
110	10	704	17.2	5.4	6.6	27.9
130	2	398	3.7	4.8	-5.7	13.2
130	4	601	10.0	5.1	-0.1	20.1
130	6	830	18.7	6.9	5.2	32.2
130	8	983	25.5	2.5	20.6	30.4

Figures

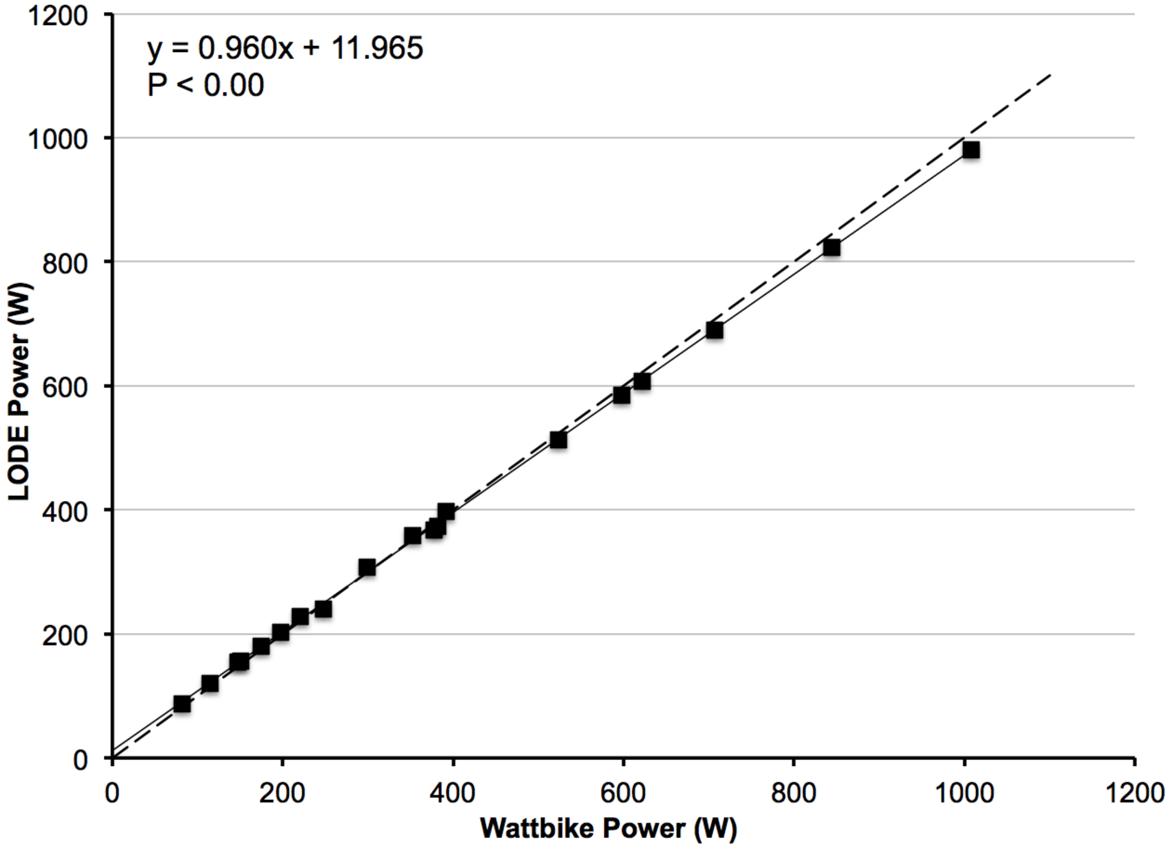

Figure 1. Scatterplot, regression model and line of identity showing the relationship between Wattbike power and LODE power for Bike 5.

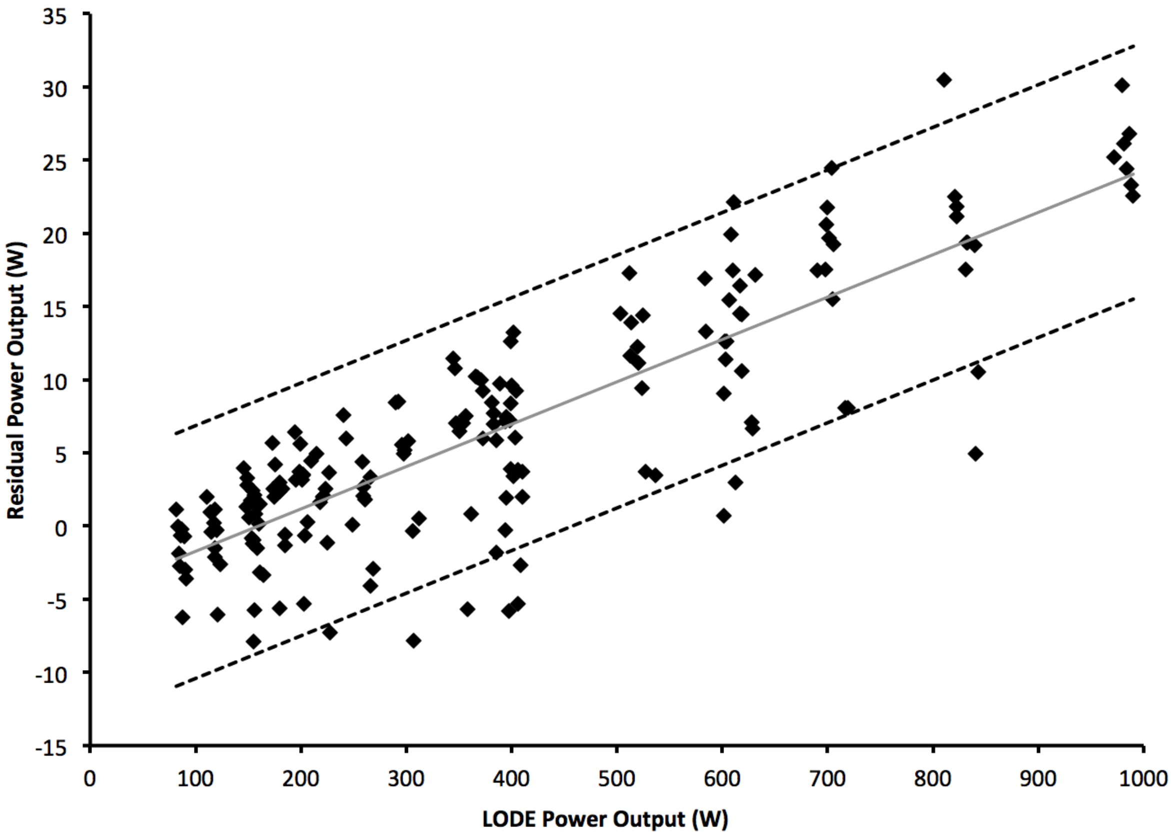

Figure 2. Scatterplot to show the differences between the LODE and Wattbikes across the range of power outputs. The 95% Limits of Agreement (dashed lines) and mean bias (solid line) represent a regression model that was calculated according to the methods of Bland & Altman (1999) for heteroscedastic data.

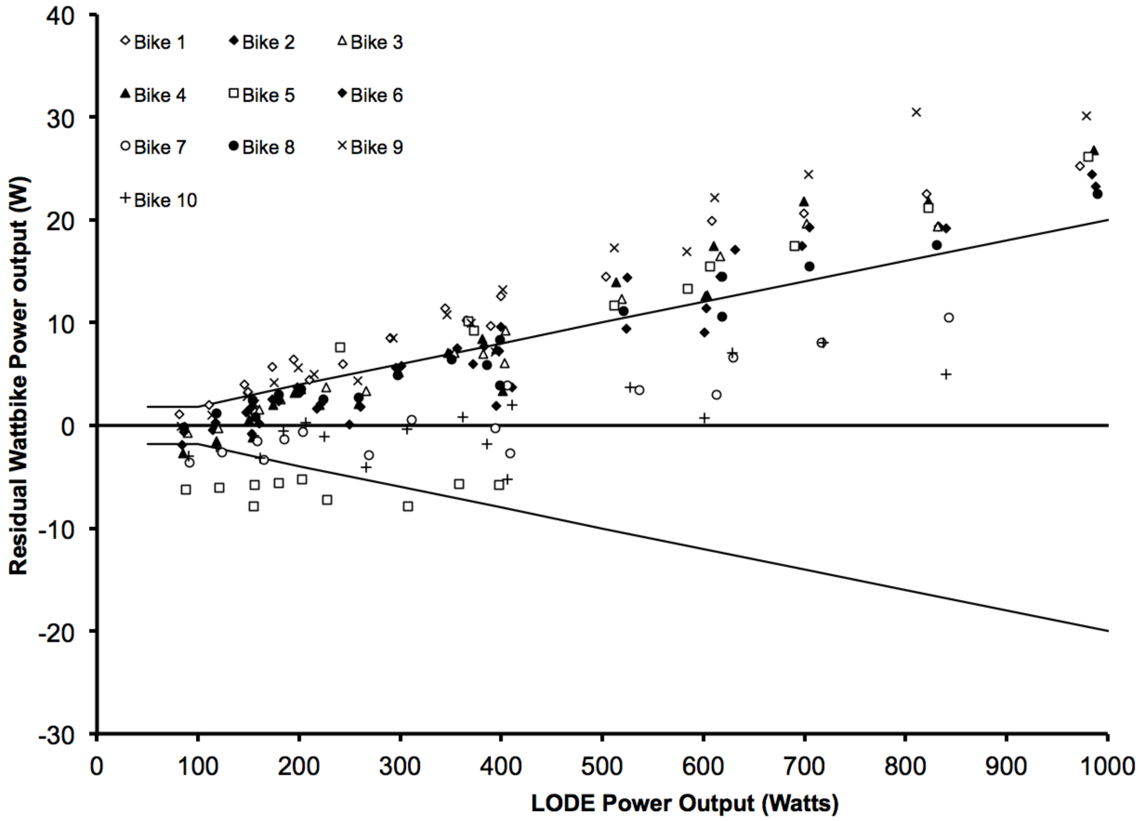

Figure 3. Scatterplot to show the differences between the LODE and Wattbikes across the range of power outputs. The black lines represent the LODE measurement error of ±2%.

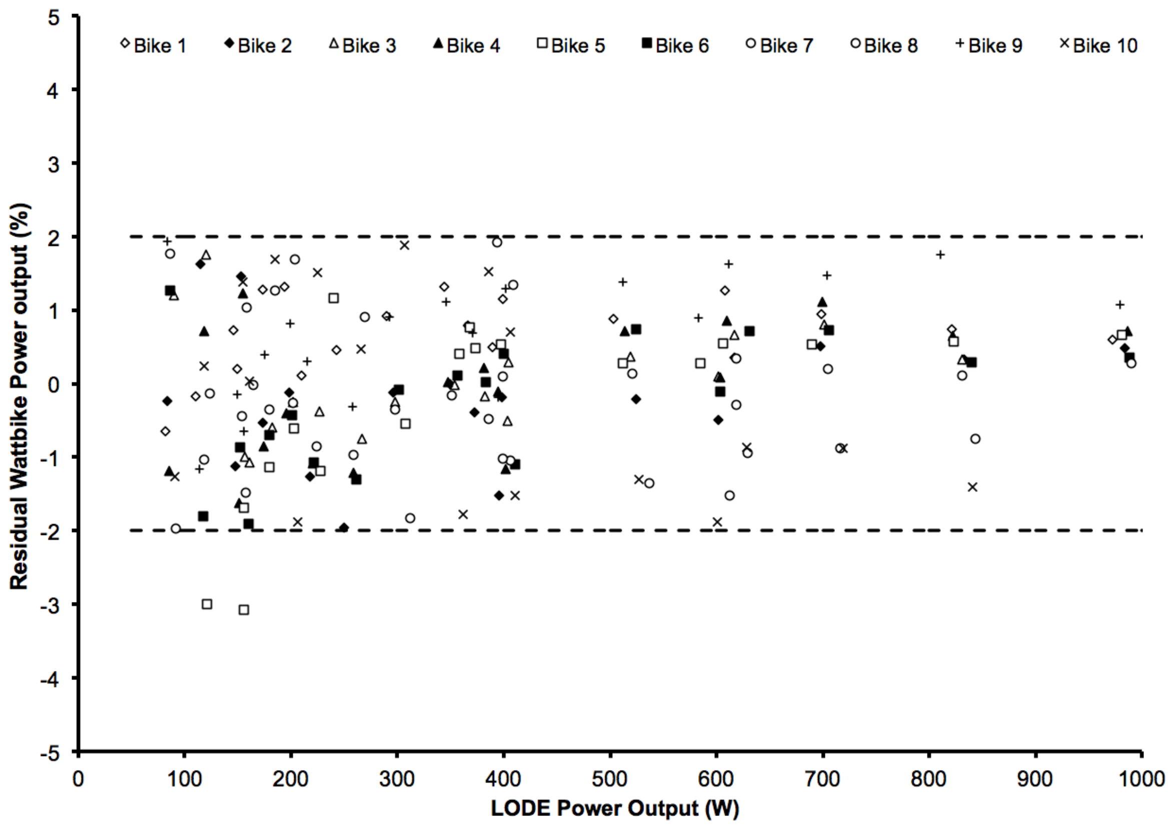

Figure 4. Scatterplot to show the percentage differences between the Wattbike and the LODE at either +2% or -2% of the LODE value.

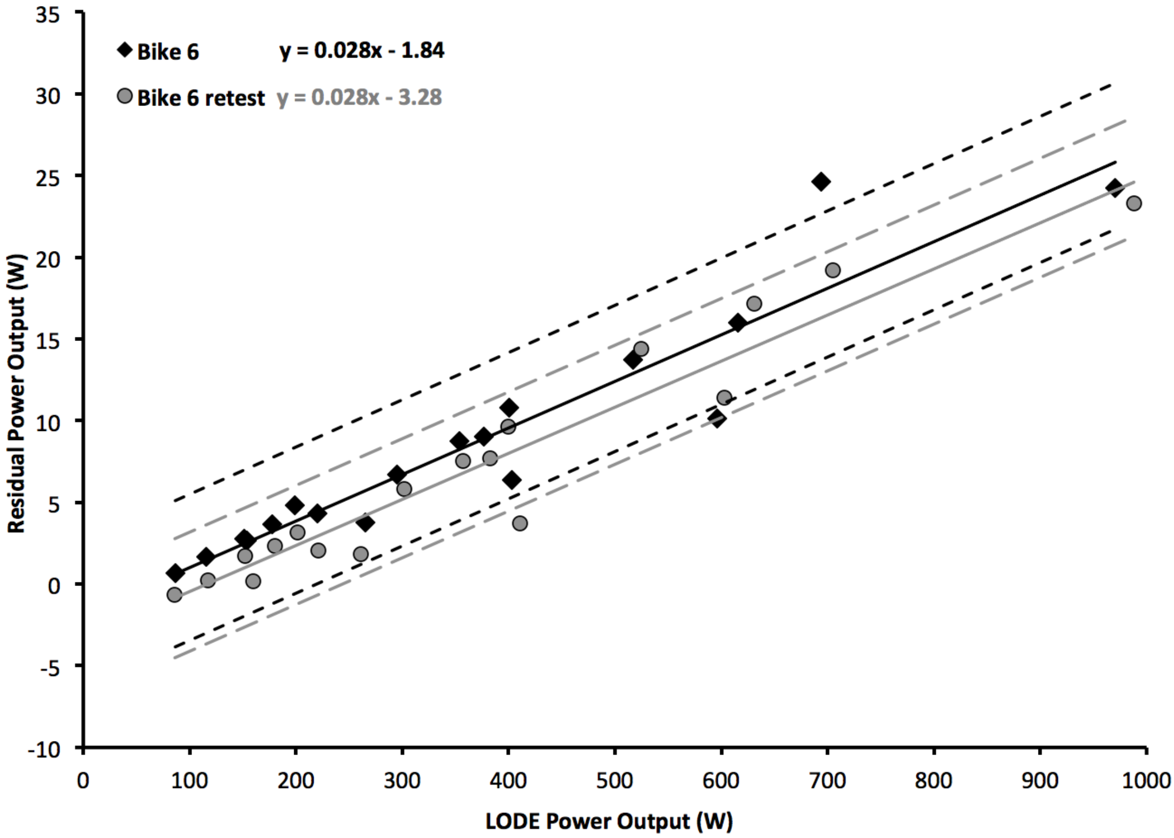

Figure 5. Scatterplot showing the similarity of the data from Bike 6 following a repeated test. The 95% Limits of Agreement (dashed lines) and mean bias (solid line) represent regression models that were calculated according to the methods of Bland & Altman (1999) for heteroscedastic data.